*Nucleic Acids Research, 2006, Vol. 34, Database issue* ***D781–D784***
*doi:10.1093/nar/gkj088*

# IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences

**Véronique Giudicelli[1], Patrice Duroux[1], Chantal Ginestoux[1], Géraldine Folch[1], Joumana Jabado-Michaloud[1], Denys Chaume[1] and Marie-Paule Lefranc[1,2,*]**

[1]IMGT®, The International ImMunoGeneTics Information System®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, Institut de Génétique Humaine IGH, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France and [2]Institut Universitaire de France, 103 Boulevard St Michel, 75005 Paris, France

Received July 26, 2005; Revised and Accepted October 13, 2005

## ABSTRACT

**IMGT/LIGM-DB is the IMGT® comprehensive database of immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences from human and other vertebrate species. It was created in 1989 by LIGM, Montpellier, France and is the oldest and the largest database of IMGT®. IMGT/LIGM-DB includes all germline (non-rearranged) and rearranged IG and TR genomic DNA (gDNA) and complementary DNA (cDNA) sequences published in generalist databases. IMGT/LIGM-DB allows searches from the Web interface according to biological and immunogenetic criteria through five distinct modules depending on the user interest. For a given entry, nine types of display are available including the IMGT flat file, the translation of the coding regions and the analysis by the IMGT/V-QUEST tool. IMGT/LIGM-DB distributes expertly annotated sequences. The annotations hugely enhance the quality and the accuracy of the distributed detailed information. They include the sequence identification, the gene and allele classification, the constitutive and specific motif description, the codon and amino acid numbering, and the sequence obtaining information, according to the main concepts of IMGT-ONTOLOGY. They represent the main source of IG and TR gene and allele knowledge stored in IMGT/GENE-DB and in the IMGT reference directory. IMGT/LIGM-DB is freely available at http://imgt.cines.fr.**

## INTRODUCTION

IMGT/LIGM-DB is the comprehensive IMGT® database of immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences from human and other vertebrate species, created in 1989 by Marie-Paule Lefranc, LIGM, Montpellier, France, on the Web since July 1995 (1–3). IMGT/LIGM-DB is the first and the largest database of IMGT®, the international ImMunoGeneTics information system® (4,5). It provides standardized and detailed immunogenetics annotations.

Owing to the complexity of the IG and TR molecular genetics (6,7) that is unique to the vertebrate genomes, IMGT/LIGM-DB has to deal with (i) large germline (non-rearranged) genomic DNA (gDNA) sequences, which may involve a complete locus from several hundred kilobases to one (or more) megabase(s); (ii) rearranged gDNA sequences resulting from the recombination of V (variable), D (diversity) and J (joining) genes (V-J genes and V-D-J genes); and (iii) rearranged V-J-C (constant) and V-D-J-C complementary DNA (cDNA designated as 'mRNA' in generalist databases) sequences. The complexity is further enhanced by the characteristics of the loci and chain types in the different species (reviewed in the IMGT Repertoire) and by the mechanisms of diversity such as combinatorial diversity, N diversity, somatic hypermutation and gene conversion (6,7). Thus, the detailed sequence annotation is a huge and complex task which requires the interpretation of DNA rearrangements and recombination, of sequence polymorphisms, of nucleotide deletions and insertions at the V-J and V-D-J junctions and, for IG, of somatic hypermutations (6,7). Annotations rely on the accuracy and the coherence of IMGT-ONTOLOGY (8), the first ontology in the field of immunogenetics which has allowed to set up the rules for standardized sequence

*To whom correspondence should be addressed at IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr), Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, IGH, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel: +33 4 99 61 99 65; Fax: +33 4 99 61 99 01; Email: lefranc@ligm.igh.cnrs.fr

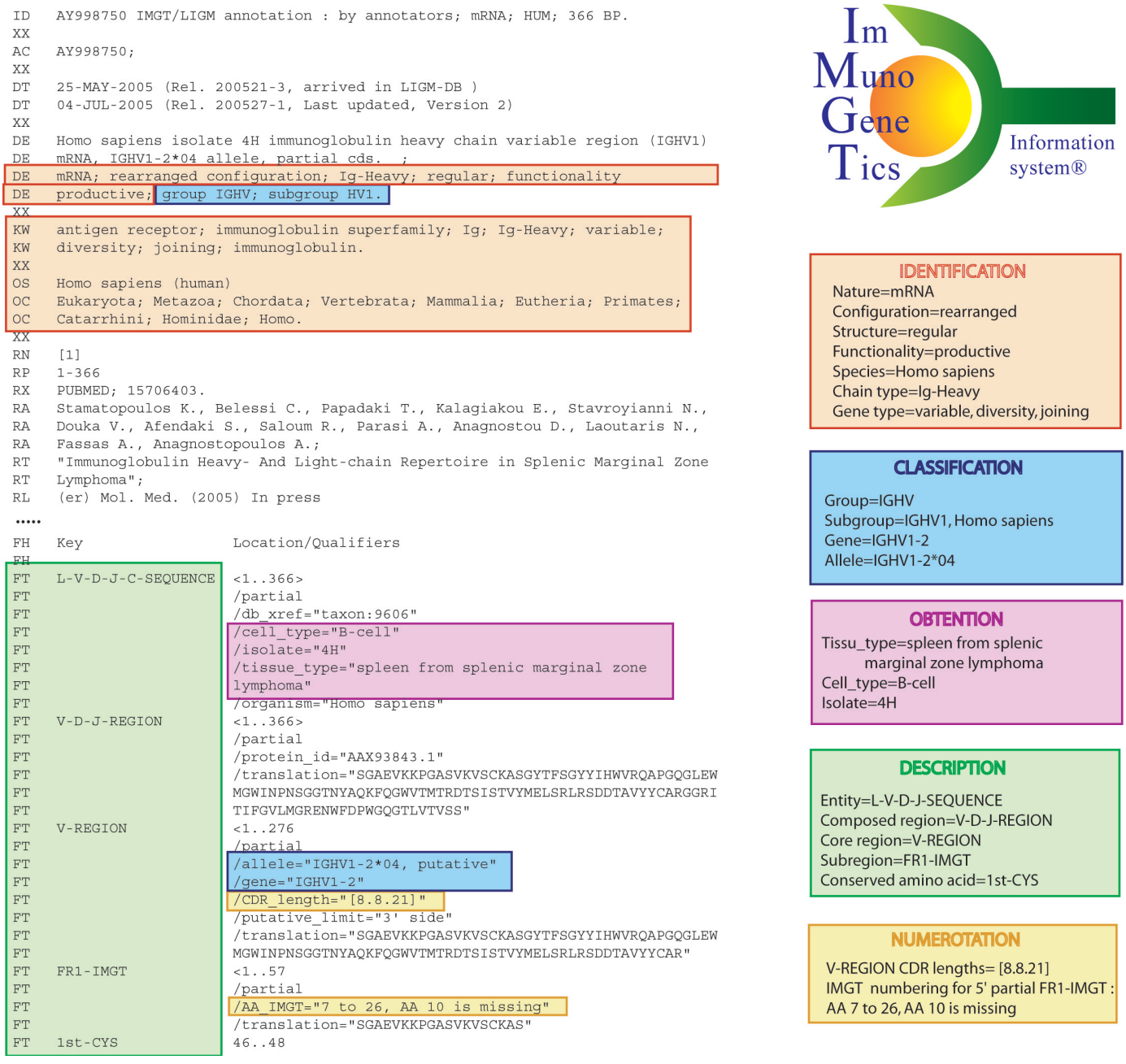


**Figure 1.** Part of a fully annotated IMGT/LIGM-DB entry according to the IMGT Scientific chart rules (5,9). The corresponding five main concepts of IMGT-ONTOLOGY (8) have been added on the right-hand side.

identification (9), gene and allele classification (6,7), constitutive and specific motif description, amino acid numbering (10–13) and sequence obtaining information.

## IMGT/LIGM-DB DATA SOURCE AND CONTENT

The unique source of IMGT/LIGM-DB nucleotide sequences is EMBL (14). Prior to being entered in IMGT/LIGM-DB, IG and TR sequences must be submitted to EMBL, GenBank or DDBJ, in order to get a unique accession number which is also the entry identifier in IMGT/LIGM-DB. Then, EMBL automatically sends the IG and TR sequences (new entries and updates) to LIGM. Sequences belonging to the human (HUM), mouse (MUS), primate (PRI), other mammals (MAM) and vertebrate (VRT) divisions, which are sufficiently reliable, are managed in IMGT/LIGM-DB, plus IG and TR-related sequences from synthetic (SYN) and unclassified (UNC). The sequences from the other EMBL divisions (CON, GSS, HTG, HTC, STS and EST) are not included. The new sequences and updates received at LIGM represent >700 sequences a week. In November 2005, IMGT/LIGM-DB

contains 98 800 sequences from 150 vertebrate species. They comprise germline gDNA, rearranged gDNA, a few germline cDNA and, for the half of the database content, rearranged cDNA (or 'mRNA'). Almost three quarters of the sequences are from human and mouse.

## IMGT/LIGM-DB ANNOTATIONS

At the reception at LIGM, data are checked by LIGM curators for their relevance. Data are then scanned to store sequences, bibliographical references and taxonomic data, whereas standardized IMGT/LIGM-DB keywords are assigned mainly manually. Based on expert analysis, specific detailed annotations are added in a second step. They follow the concepts of IMGT-ONTOLOGY (8) and the rules of the IMGT Scientific chart (9). This allows, for example for the sequence shown in Figure 1, the precise sequence identification with the characterization of the nature of the molecule, the configuration, the structure, the functionality, the species, the chain type and the gene type (IDENTIFICATION concept), the characterization of the group and subgroup, and the classification of the gene and allele according to the IMGT nomenclature (CLASSIFICATION concept) (15), the description of the constitutive immunogenetics specific motifs (DESCRIPTION concept), the codon and amino acid numbering (NUMEROTATION concept), and the sequence obtaining information (OBTENTION concept, currently in development, with an important analysis devoted to the biological origin of the sequence, the clinical specification and the description of used methodology). Most of the annotations are manually performed with the help of IMGT® tools, IMGT/V-QUEST (16) and IMGT/JunctionAnalysis (17). However, a part of human and mouse cDNA sequences have been automatically annotated by the internal tool IMGT/Automat (18,19).

## IMGT/LIGM-DB SEARCH AND DISPLAY

The IMGT/LIGM-DB data are provided with a user-friendly interface. The Web interface allows searches according to immunogenetic-specific criteria and is easy to use without any knowledge in a computing language. The interface allows the users to get easily connected from any type of platform using free browsers. All IMGT/LIGM-DB information is available through five modules of search: Catalogue, Taxonomy and Characteristics, Keywords, Annotation labels and References. Selection is displayed at the top of the 'results of your search' page, so the users can check their own queries (20). Users have the possibility to modify their request or to consult the results. They can (i) add new conditions to increase or decrease the number of resulting sequences; (ii) view details: selecting this 'View' option provides a list of resulting sequences; selection of one sequence in the list offers nine possibilities: annotations, IMGT flat file, coding regions with protein translation, catalogue and external references, sequence in dump format, sequence in FASTA format, sequence with three reading frames, EMBL flat file, IMGT/V-QUEST (16); or (iii) search for sequence fragments: selecting this 'Subsequences' option allows to search for sequence fragments (subsequences) corresponding to a particular label for the resulting sequences (available for fully annotated sequences) (20).

## IMGT/LIGM-DB DISTRIBUTION

IMGT/LIGM-DB flatfiles are available by anonymous FTP servers at CINES (ftp://ftp.cines.fr/IMGT/), at EBI (ftp://ftp.ebi.ac.uk/pub/databases/imgt/), and at IGH (ftp://ftp.igh.cnrs.fr/pub/IMGT/) and from many SRS (Sequence Retrieval System) sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (e.g. CINES, EBI, INFOBIOGEN and Institut Pasteur). IMGT/LIGM-DB data can also be retrieved through Web services which are developed and implemented with Axis (5). For instance, they include the 'queryKnowledge' which provides the lists of instances for the IMGT-ONTOLOGY concepts, and the 'querySeqData' which allows the retrieval of any sequence-related data, identified, classified, described according to the IMGT® concepts, such as the nucleotide sequence, the description labels, the literature references, the metadata, etc. The result is then a list of data entries, in IMGT-ML format, sharing these given values (5). IMGT/LIGM-DB data are cross-referenced in the EMBL databank (14), in IMGT/GENE-DB (15) which allows to link gene entries with the corresponding genomic reference sequences and with the known expressed cDNAs, and in IMGT/PRIMER-DB (21) in order to display the oligonucleotide primers within the sequences.

## CONCLUSION AND PERSPECTIVES

IMGT/LIGM-DB manages all published vertebrate IG and TR nucleotide sequences. Very interestingly and despite the complexity of these sequences, and their variability in many species (sequences of 150 species are dealt in IMGT/LIGM-DB), the detailed annotations are all performed according to the concepts of IMGT-ONTOLOGY. The organization of the concepts has been formalized, with XML Schema, in IMGT-ML (5). A new IMGT/LIGM-DB interface, available in 2006, will allow queries according to these concepts and the retrieval of entries as XML files, in IMGT-ML format.

## CITATION

Users of IMGT/LIGM-DB are requested to cite this article in their publications and to quote the IMGT® home page URL (http://imgt.cines.fr).

## ACKNOWLEDGEMENTS

We thank Oliver Clément and Gérard Lefranc for helpful discussion. We are deeply grateful to the IMGT® team for its expertise and constant motivation and especially to our curators for their hard work and enthusiasm. IMGT® was funded by the European Union's BIOMED2, BIOTECH, 5th PCRDT programme (QLG2-2000-01287). IMGT® has obtained the Plate-forme RIO label since 2001. IMGT® is a CNRS trademark. IMGT® is currently funded by the Centre National de la Recherche Scientifique (CNRS) and the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la

Recherche (MENESR) (Université Montpellier II Plan-Pluri-Formation, BioSTIC-LR2004, ACI-IMPBIO IMP82-2004, GIS AGENAE and Réseau National des Genopoles RNG). Funding to pay the Open Access publication charges for this article was provided by CNRS.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Lefranc,M.-P., Giudicelli,V., Busin,C., Malik,A., Mougenot,I., Déhais,P. and Chaume,D. (1995) LIGM-DB/IMGT: an integrated database of Ig and TcR, part of the immunogenetics database. *Annal N. Y. Acad. Sci*., **764**, 47–49.
2. Giudicelli,V., Chaume,D., Bodmer,J., Müller,W., Busin,C., Marsh,S., Bontrop,R., Lemaître,M., Malik,A. and Lefranc,M.-P. (1997) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res*., **25**, 206–211.
3. Lefranc,M.-P. (2004) IMGT, the international ImMunoGenetics information system®, http://imgt.cines.fr. In Lo,B.K.C. (ed.), *Antibody Engineering Methods and Protocols*, *2nd edn*, *Methods in Molecular Biology,* Humana Press, Totowa, NJ, Vol. 248, Ch. 3, pp. 27–49.
4. Lefranc,M.-P., Giudicelli,V., Kaas,Q., Duprat,E., Jabado-Michaloud,J., Scaviner,D., Ginestoux,C., Clément,O., Chaume,D. and Lefranc,G. (2005) IMGT, the international ImMunoGeneTics information system®. *Nucleic Acids Res*., **33**, D593–D597.
5. Lefranc,M.-P., Clément,O., Kaas,Q., Duprat,E., Chastellan,P., Coelho,I., Combres,K., Ginestoux,C., Giudicelli,V., Chaume,D. *et al*. (2005) IMGT-Choreography for Immunogenetics and Immunoinformatics. *In Silico Biol*., **5**, 45–60.
6. Lefranc,M.-P. and Lefranc,G. (2001) *The Immunoglobulin FactsBook*. Academic Press, London, UK, 458 pages.
7. Lefranc,M.-P. and Lefranc,G. (2001) *The T cell receptor FactsBook*. Academic Press, London, UK, 398 pages.
8. Giudicelli,V. and Lefranc,M.-P. (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics*, **15**, 1047–1054.
9. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Bosc,N., Folch,G., Guiraudou,D., Jabado-Michaloud,J., Magris,S., Scaviner,D., Thouvenin,V. *et al*. (2004) IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol*., **4**, 17–29.
10. Lefranc,M.-P. (1997) Unique database numbering system for immunogenetic analysis. *Immunol. Today*, **18**, 509.
11. Lefranc,M.-P. (1999) The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. *Immunologist*, **7**, 132–136.
12. Lefranc,M.-P., Pommié,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol*., **27**, 55–77.
13. Lefranc,M.P., Pommié,C., Kaas,Q., Duprat,E., Bosc,N., Guiraudou,D., Jean,C., Ruiz,M., Da Piedade,I., Rouard,M. *et al*. (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol*., **29**, 185–203.
14. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al*. (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res*., **33**, D29–D33.
15. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res*., **33**, D256–D261.
16. Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res*., **32**, W435–W440.
17. Yousfi Monod,M., Giudicelli,V., Chaume,D. and Lefranc,M.-P. (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20**, I379–I385.
18. Giudicelli,V., Protat,C. and Lefranc,M.-P. (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In *Proceedings of the European Conference on Computational Biology ECCB'2003*, September 27–30, Paris, France, PS-32, DKB-31, pp. 103–104.
19. Giudicelli,V., Chaume,D., Jabado-Michaloud,J. and Lefranc,M.-P. (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud. Health Technol. Inform*., **116**, 3–8.
20. Lefranc,M.-P., Giudicelli,V., Ginestoux,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbié,V. and Chaume,D. (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res*., **27**, 209–212.
21. Folch,G., Bertrand,J., Lemaitre,M. and Lefranc,M.-P. (2004) IMGT/PRIMER-DB. In Galperin M.Y. (ed.). Database listing, The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res*., **32**, D3–D22.